Math 455 Class 16 February 25

Reminder: no class Friday Feb 27.

Midterm 1: Monday, March 2, in class

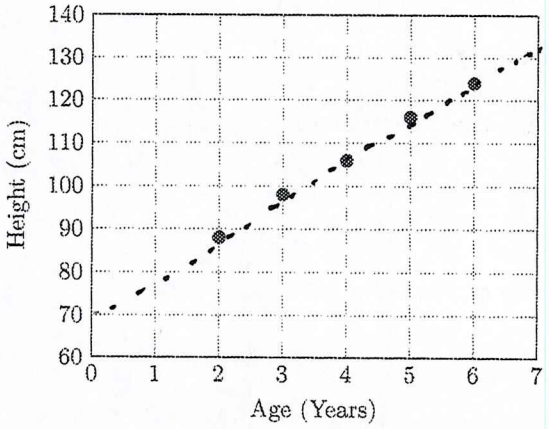
no books notes or calculators.

Linear Regression: Estimating the Growth Curve

Scenario: A researcher is tracking the height of a child named Jordan to see how well a linear model predicts future growth. The following measurements were taken between the ages of 2 and 6.

Jordan's Growth Data

Age (x)	Height (y)
2	88 cm
3	98 cm
4	106 cm
5	116 cm
6	124 cm



Tasks

1. **Visual Estimation:** Use the table and the graph to estimate the following:
   - **The Slope (m):** Based on the annual changes in the table, what is the average growth rate in cm per year?
   - **The y-intercept (b):** Based on the data and graph, what is the estimated intercept b?
2. **The Linear Model:** Write your estimated equation for the line of best fit:

$$\hat{y} = mx + b$$

3. **Interpretation:** Write a sentence interpreting the y-intercept b. What are the units in which b is measured?
4. **Prediction:** Use your equation to predict Jordan's measured height at **Age 4.5**.
5. **Future Prediction:** Use your equation to predict Jordan's height at **Age 21**.
6. **Context:** Do you believe the predictions? Why or why not?

Fitting a linear model:

$$\text{Height} = b + m\,\text{Age} + \varepsilon$$

What are $\hat{b}$ and $\hat{m}$? (approx.)

Linear regression is drawing a line through a bunch of points. Our points are almost on a line already!

We have the data and we COULD compute

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

and similarly $\hat{\beta}_0$.

$\hat{\beta}_1 = \hat{m}$

$\hat{\beta}_0 = \hat{b}$.

We don't HAVE to do this. The question asks for an approximate answer: since the points are almost in a line, draw the line.

We SEE $\hat{b} \approx 70$ and $\hat{m} \approx 9$.

If you insist on calculating something:
draw a line through the points. (2, 88) and (6, 124)

this has slope $\frac{124-88}{6-2} = \frac{36}{4} = 9$.

So our "visual" regression line is

$$\widehat{\text{Height}} = 70 + 9 \cdot \text{Age}$$

(3) The interpretation of $\hat{b}$:

The child's height (length?) at birth is 70 cm.

(Note: measured in cm.)

(4) Height at age 4.5 $= 70 + 9 \cdot (4.5) = 110.5$ cm

(5) Height at age 21. $= 70 + 9 \cdot 21 = 259$ cm

(6) The prediction for age 4.5 is reasonable.

The prediction for age 21 is <u>NOT</u>:

Jordan is predicted to be 8+ feet tall.

We have "domain-specific knowledge": people don't grow 9cm per year between the ages of (say) 17 and 21.

These approximations come up in the theory of the Bootstrap (Ch. 3) and the Bonferroni correction. (Ch. 6)

Q: What is the Bonferroni correction?

A: Imagine that we are trying to get our article in [Social Science of your choice] published and we need a p-value < 0.05.

We have the following idea: if we need a p-value less than 0.05, let's just test 20 things!

Even if H_0 is true, one of them will (on average) have a p-value less than 0.05.

Of course, doing so would be less than honest, especially if we conceal the 19 tests that failed.

The Bonferroni Correction is a way of accounting for this.

Linear regression (or your favorite ML technique) does not know this!

Q: Is this due to the uncertainty being larger far from the mean of the data?

A: No. It is because the linear model is WRONG outside of a small range of ages.

- Domain-specific knowledge is important and we may override a model if we think it is wrong.
- Never forget that linear regression is drawing a line: there is no magic.

An approximation from Calc 2:

For large n $\left(1+\frac{x}{n}\right)^n \approx e^x$.

in particular: $\left(1-\frac{1}{n}\right)^n \approx \frac{1}{e}$

For small x: $e^x \approx 1+x$.

It says that if we perform n tests at level α, we should replace α by $\frac{\alpha}{n}$ to be confident that our result is "real." (i.e. not just luck from performing so many tests.)

$$P(\text{all tests accept } H_0) = 1 - P(\text{at least one rejects})$$

$$\geq 1 - \sum_{i=1}^{n} P(\text{test } i \text{ rejects } H_0) = 1 - n\alpha.$$

This says that to "get" $1-\alpha$ we need to replace α by $\frac{\alpha}{n}$ so $1 - n\left(\frac{\alpha}{n}\right) = 1-\alpha$.

Or: $\left(1 - \frac{\alpha}{n}\right)^n \simeq e^{-\alpha} \simeq 1-\alpha$.

Text book material:

Wackerly Ch. 7, 10.

Faraway Ch. 1-5.

Note Faraway Ch. 5 includes Hill Criteria.